BankIt: A Tool for Simple GenBank Submission

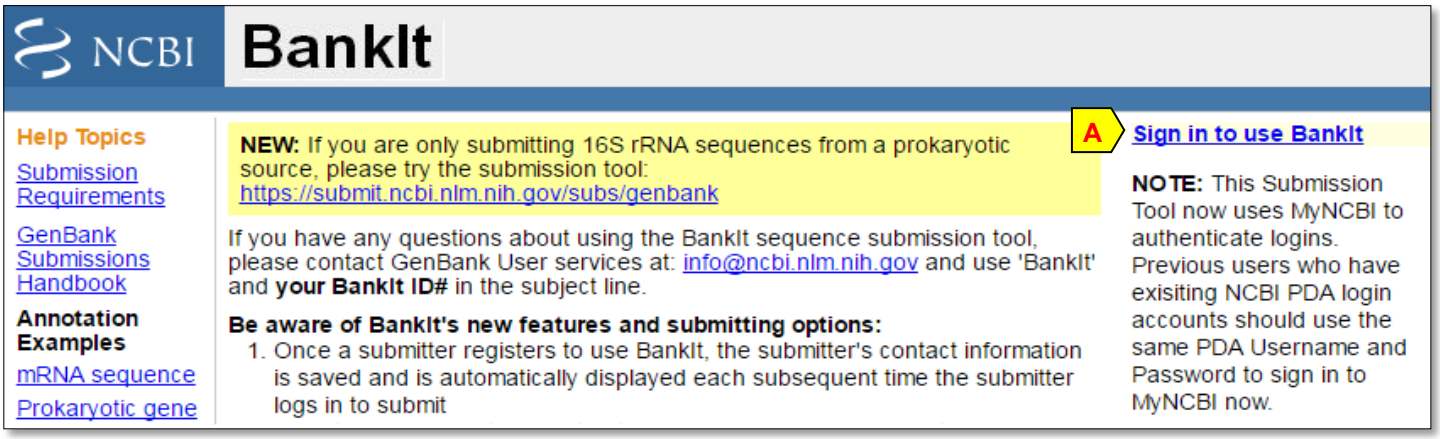
https://submit.ncbi.nlm.nih.gov/WebSub/?tool=genbank/

Submitting a single sequence or small batch of different sequences or a simple set of sequences

National Center for Biotechnology Information • National Library of Medicine • National Institutes of Health • Department of Health and Human Services

Introduction

BankIt is a web-based tool for submission of primary sequences or sequences from Third Party Annotation (TPA) efforts to GenBank. If you are a first-time submitter, please review the GenBank Submission Policy (www.ncbi.nlm.nih.gov/WebSub/?tool=genbank) to see the sequence types accepted by GenBank and familiarize yourself with the field definitions of a typical GenBank record (www.ncbi.nlm.nih.gov/Sitemap/samplerecord.html). Before starting a BankIt submission, gather enough information to meet the requirements for GenBank submissions through BankIt, as listed in this document www.ncbi.nlm.nih.gov/WebSub/html/requirements.html. BankIt now uses MyNCBI to authenticate logins. Go to the BankIt homepage (www.ncbi.nlm.nih.gov/WebSub/?tool=genbank), and click on "Sign in to use BankIt" (**A**) to log in. If you don't have an MyNCBI account, click on "Register for an NCBI account" in the login page to create one. More details are at www.ncbi.nlm.nih.gov/books/BK3842/#MyNCBI.Registering_with_My_NCBI.



Steps to complete a BankIt submission

GenBank issues a unique BankIt number (**B**) for each new submission, which helps identify your submission and track the sequences you have submitted. A BankIt submission contains a set of distinctive steps (listed as tabs at the top, **C**), each with its own form for you to input the required data. To complete a submission using BankIt, you can work through the pages sequentially and enter essential information for each and every form. Once you have provided all the information correctly and clicked the "Continue" button, you will go to the next step. If necessary, you can go back to make revisions, or stop at any step and continue at a later time.

At the **Contact** step, provide an alternative email and check the option to save the contact information for multiple or future BankIt submissions; at the **Reference** step, if no published paper is available, enter a tentative reference title describing the research from which the submitted sequences were obtained, you can update it after the publication; at the **Sequencing Technology** step, for sequences assembled from raw reads generated by Next Generation Sequencing technologies (NGS), you need to select the option of assembled sequences (consisting of two or more sequence reads) and provide relevant assembly information; at the **Nucleotide** step, provide a release date if you want your sequences to be confidential until published (HUP), then enter the number of nucleotide sequences in your current submission, and paste the plain text FASTA sequences in the box, or upload a FASTA sequence file prepared as described (www.ncbi.nlm.nih.gov/WebSub/html/help/fasta.html); at the **Set/Batch** step, choose Batch if your sequences are from different molecule types or genes, or select an appropriate set for the sequences from the same gene or genomic region but from different species or samples.

Steps to complete a BankIt submission (cont.)

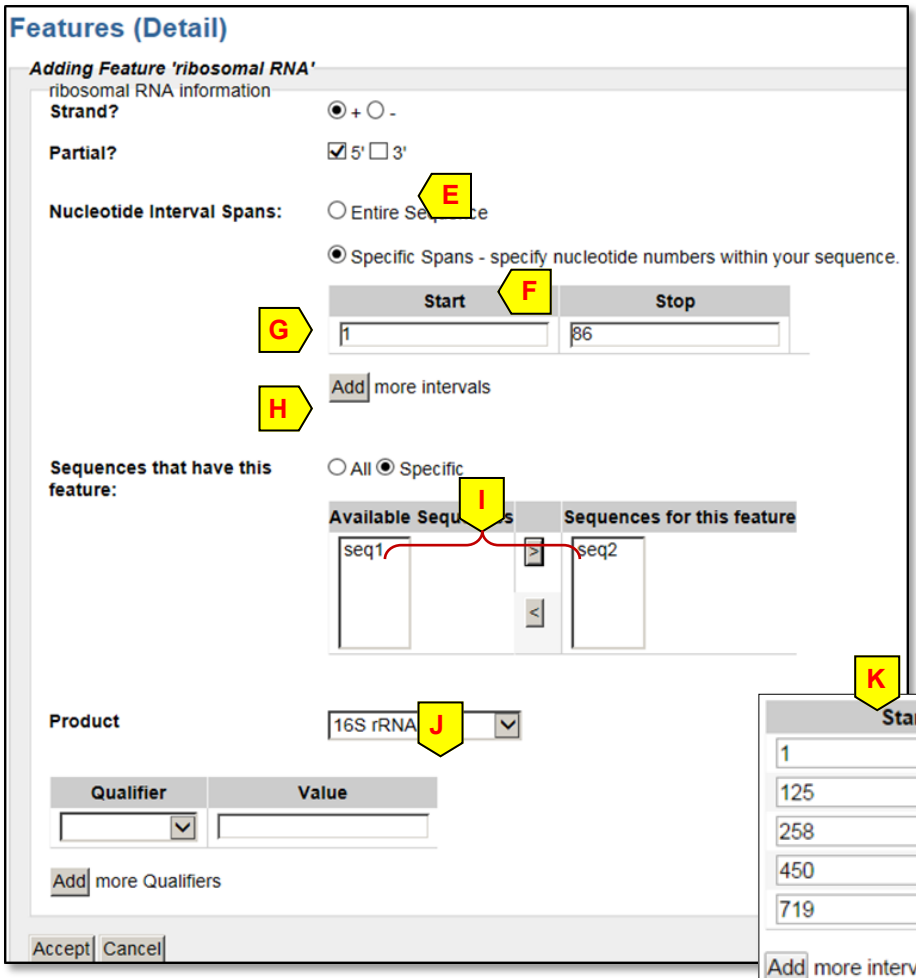
At the **Organism** step, you will need to provide the name of the organism from which the sequence was isolated, if your FASTA sequences' definition lines do not encode organism name; at the **Submission Category** step, please check for what should not be submitted if your submission is TPA in nature (www.ncbi.nlm.nih.gov/genbank/tpa-exp & www.ncbi.nlm.nih.gov/genbank/tpa-inf); at the **Source Modifiers** step, select the relevant organelle, as BankIt needs this information to apply the correct genetic code for the CDS annotation. Use valid names and formats of source modifiers (listed at www.ncbi.nlm.nih.gov/WebSub/html/help/genbank-source-table.html) and approved names of countries (listed at www.ncbi.nlm.nih.gov/genbank/collab/country/), avoid non-ASCII characters and symbols all together.

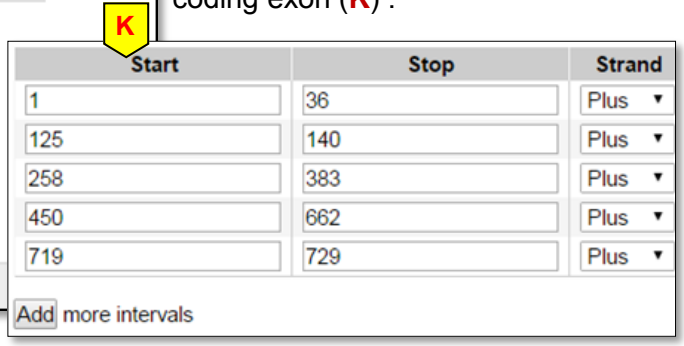
Sequence_ID	Collected_by (A)	Collection_date (B)	Country (C)	Lat_Lon (D)
Sequence108	James Schandón	16-sep-11	Columbia	4°45' N 74°28' W

The example source table contains several formatting issues: **A**) a non-ascii character, **B**) an incorrect format of collection_date, **C**) a misspelling in the country name, and **D**) an incorrect format of Lat_Lon with symbols. The correctly formatted source modifiers table is shown below.

Sequence_ID	Collected_by	Collection_date	Country	Lat_Lon
Sequence108	James Schandon	16-Sep-2011	Colombia	4.75 N 74.47 W

At the **Features** step, you can annotate your sequences with appropriate features by either filling up the forms or uploading a feature table file. If you choose to use forms, you need to provide detailed information about the features you selected, including the strand (**E**) on which the feature appears, whether the feature is partial (**F**), and whether the nucleotide interval where the feature occurs spans the entire sequence or has a specific span (**G**). For a specific span, you must provide the sequence coordinates (nucleotide numbers, **H**) of that span. When submitting multiple sequences in the same batch, you need to indicate whether all sequences or specific sequences have this feature by selecting them from the sequence list (**I**) and annotate them together in the same batch. Please note that even if sequences in a multi-sequence submission have the same feature (**J**), the reading frames or specific spans may be different and/or located on a different strand. Therefore, you may still need to annotate them separately. For gene sequences with multiple coding exons, you need to enter intervals for each and every coding exon (**K**) .

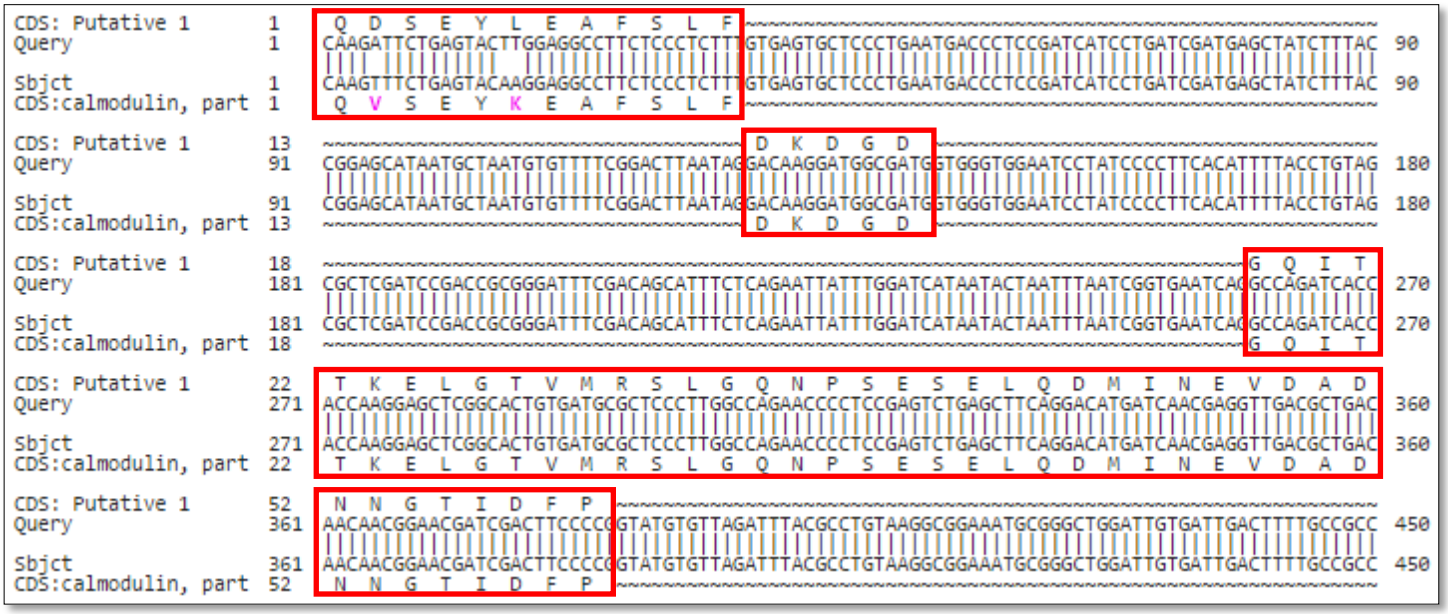




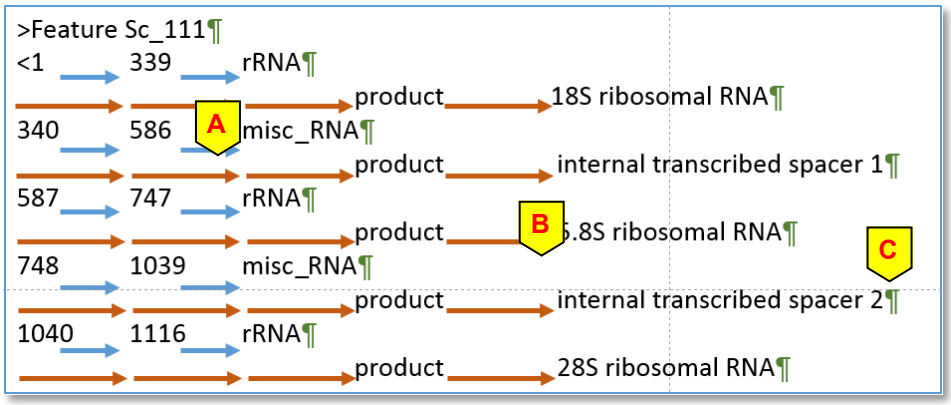
Steps to complete a BankIt submission (cont.)

Most eukaryotic genes, such as GAPDH and ACTB, have multiple coding exons. Instead of inputting specific spans for each exon interval, you can choose to upload a protein sequence. This protein sequence ID should match its corresponding nucleotide sequence ID, and its length should not exceed the maximal length of a protein that could possibly be encoded by the nucleotide sequence. For example, a nucleotide sequence 200 bases in length can encode a protein sequence maximally with no more than 66 amino acid residues.

You can use the BLAST service from NCBI to help you find the encoded protein sequence. To do so, run a nucleotide BLAST search (http://1.usa.gov/1XHpcrk) with your nucleotide sequence. Once you get the BLAST hits, click the "Reformatting options," check the "CDS feature" checkbox, and press the "Reformat" button to show the CDS region in the alignment. By matching your sequence against those with CDS annotations, you will be able to get the encoded protein sequence from multiple coding exonic regions, as outlined in red. Note that GenBank requires submitters to provide a complete name of the translated protein product while annotating a CDS feature.



In addition to using online BankIt forms to annotate your sequences, you may create a five-column, tab-delimited feature table as described at www.ncbi.nlm.nih.gov/Sequin/table.html. In order to make the feature table importable by BankIt, the SeqID in your feature table must match with the SeqID in the corresponding FASTA nucleotide sequence file. For example, if you have a nucleotide SeqID such as Sc_111, shown in the definition line as ">**Sc_111** [organism=Saccharomyces cerevisiae]," the SeqID in the feature table must be:



">Feature **Sc_111**" (both without quotes). It is critical to format a feature table correctly. Otherwise, BankIt will not be able to import it correctly, or import it at all. Note that in a feature table, there are five columns separated only by a single tab (**A, B**) between columns on two lines. There should be no redundant blank spaces or tabs between columns, or extra carriage returns between lines (**C**).

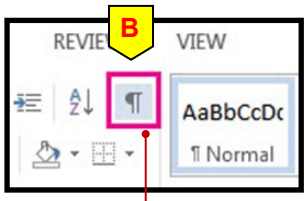
Note: A feature is defined by two or more lines. Start and end coordinates always go to the first 2 columns, with the third in the first coordinate line specifying the feature. Feature qualifier and value (http://www.insdc.org/feature_table.html#3.3) go to columns 4 and 5 in separate lines below the lines for coordinates, with their first 3 columns left blank.

Steps to complete a BankIt submission (cont.)

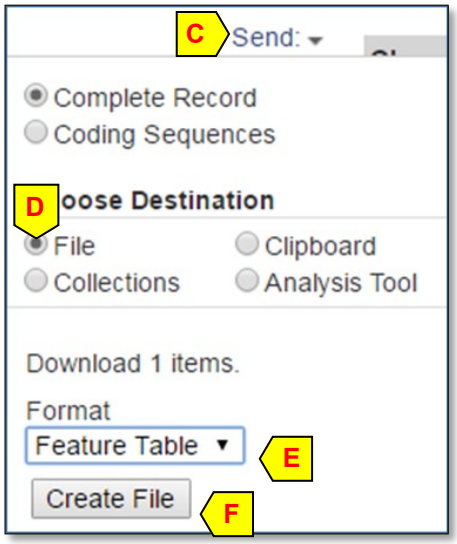
You can use a spreadsheet, such as Microsoft Excel, to help track the number of tabs in the feature table you are constructing, by entering the data only in specific cells (**A**), and saving the file in the tab-delimited plain-text format. You can view those tabs in your saved feature table file within Microsoft Word by clicking the "Show/Hide" button (**B**).

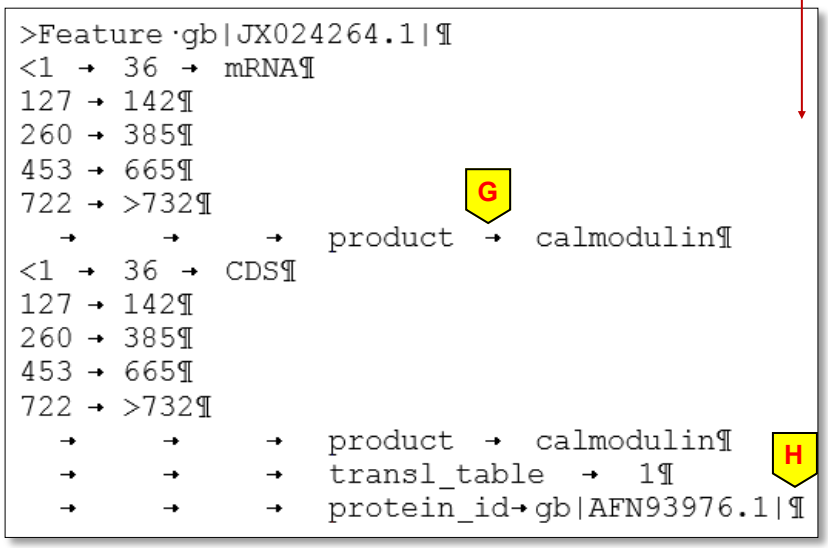
A

	A	B	C	D	E
1	>Feature Sc_111				
2	<1	>2280	gene		
3				gene	PB2
4	<1	>2280	CDS		
5				product	polymerase PB2
6				codon_start	2



You can also run BLAST search to find similar sequences in the NT BLAST database with feature annotations, then retrieve the record from the Nucleotide database by clicking its accession and download its Feature Table for use as a template. Click "Send (**C**) >> File (**D**) >> Feature Table (**E**) >> Create File (**F**)" to save the feature table to a file. As an example, to download the Feature Table for GenBank record JX024264, open the record (www.ncbi.nlm.nih.gov/nuccore/JX024264), then follow the steps above to download the table.





Above right is the downloaded table for the annotation of both mRNA and CDS features of JX024264, viewed in Microsoft Word with tabs and carriage returns displayed. The arrows in the table are tabs (**G**) and the symbol at the end are carriage returns (**H**). If you are going to use this table as a template for your sequence annotation, you need to check the alignment between your sequence and the JX024264 sequence to see if there are any insertions or deletions within the range of alignment in your sequence and adjust the nucleotide coordinates of specific spans in your table accordingly. To verify that your feature table with CDS annotation is correct, you can upload the table into BankIt, get the translated protein sequence, then align your translated protein sequence with the one from the GenBank record using the "Align two or more sequences" option of on the protein BLAST page. You will need to revise your feature table if there are significant discrepancies between the two protein sequences. After finishing feature annotation, you will reach the final **Review and Correct** step, where you have the options to provide additional information or let GenBank staff know that you are resubmitting your sequences. Note that clicking the **Finish Submission** button completes your BankIt submission, which cannot be revised. If you are uncertain or have questions about your submission, do not click that button. Instead send your questions to info@ncbi.nlm.nih.gov to have them addressed first, then edit your unfinished submission afterward. GenBank will send you an acknowledgement email for your completed submission.

Additional Information

https://www.ncbi.nlm.nih.gov/books/NBK63586/
https://www.youtube.com/watch?v=OZxxsRm0pP4
https://www.youtube.com/watch?v=DhYUYJSm2mQ